University of Washington
Department of Electrical Engineering
EE512 Spring, 2006
Graphical Models

Jeff A. Bilmes <bilmes@ee.washington.edu>

Lecture 12 Slides
May 4th, 2006 (star wars day)
Announcements

• READING:
  – M. Jordan: Chapters 4, 10, 12, 17, 18

• Reminder: TA discussions and office hours:
  – Office hours: Thursdays 3:30-4:30, Sieg Ground Floor Tutorial Center
  – Discussion Sections: Fridays 9:30-10:30, Sieg Ground Floor Tutorial Center Lecture Room

• Reminder: take-home Midterm: May 5th-12th, you must work alone on this.

• Office Hours Yesterday

• Note: I am gone all next week (May 8th-12th). Subsequent weeks we will have makeup classes on Mondays at 6:00pm.
Class Road Map

• L1: Tues, 3/28: Overview, GMs, Intro BNs.
• L2: Thur, 3/30: semantics of BNs + UGMs
• L3: Tues, 4/4: elimination, probs, chordal I
• L4: Thur, 4/6: chrdl, sep, decomp, elim
• L5: Tue, 4/11: chdl/elim, mcs, triang, ci props.
• L6: Thur, 4/13: MST,CI axioms, Markov prps.
• L7: Tues, 4/18: Mobius, HC-thm, (F)=(G)
• L8: Thur, 4/20: phylogenetic trees, HMMs
• L9: Tue, 4/25: HMMs, inference on trees
• L10: Thur, 4/27: Inference on trees, start poly

• L11: Tues, 5/2: polytrees, start JT inference
• L12: Thur, 5/4: Inference in JTs
• L13: Tues, 5/9: away
• L14: Thur, 5/11: away
• L15: Tue, 5/16
• L16: Thur, 5/18
• Mon, 5/22
• L17: Tues, 5/23
• L18: Thur, 5/25
• Mon, 5/29
• L19: Tue, 5/30
• L20: Thur, 6/1: final presentations
Final Project Milestone Due Dates

• L1: Tues, 3/28:
• L2: Thur, 3/30:
• L3: Tues, 4/4:
• L4: Thur, 4/6:
• L5: Tue, 4/11:
• L6: Thur, 4/13:
• L7: Tues, 4/18:
• L8: Thur, 4/20: Team Lists, short abstracts I
• L9: Tue, 4/25:
• L10: Thur, 4/27: short abstracts II

• L11: Tues, 5/2: today
• L12: Thur, 5/4: abstract II + progress
• L13: Tues, 5/9
• L14: Thur, 5/11: 1 page progress report
• L15: Tue, 5/16
• L16: Thur, 5/18: 1 page progress report
• L17: Tues, 5/23
• L18: Thur, 5/25: 1 page progress report
• L19: Tue, 5/30
• L20: Thur, 6/1: final presentations
• L21: Tue, 6/6 4-page papers due (like a conference paper).

• Team lists, abstracts, and progress reports must be turned in, in class and using paper (dead tree versions only).
• Final reports must be turned in electronically in PDF (no other formats accepted).
• Progress reports must report who did what so far!!
Summary of Last Time

• inference on poly trees
• Begin exact inference on junction trees
Outline of Today’s Lecture

• exact inference on junction trees
Books and Sources for Today

• M. Jordan: Chapters 4, 10, 12, 17, 18
Inference

- If starting from a BN, we can convert to a MRF by moralization.
- For each CPT, find some clique that can hold it, and assign it there (doesn’t matter which one, as long as there is only one). When this is done, we will have that:

\[
p(x) = \prod_{\nu} p(x_{\nu} | x_{pa(\nu)}) = \prod_{c \in C} \psi_c(x)
\]

- Assign evidence. We can view this as zeroing out tables in the clique potentials \( \psi_c(x_c) \) in accordance to the evidence.
- Alternatively (and equivalently), just multiply in delta functions.
Goal calculation & Evidence

\[ p(x_H | \bar{x}_E) = p(x_H, \bar{x}_E) \frac{1}{p(\bar{x}_E)} = \]

\[ = \frac{\Pi_c \psi_c(x_H, \bar{x}_E)}{Z} \frac{Z}{\sum_{x_H} \Pi_c \psi_c(x_H, \bar{x}_E)} = \frac{\Pi_c \psi_c(x_H, \bar{x}_E)}{\sum_{x_H} \Pi_c \psi_c(x_H, \bar{x}_E)} \]

- Therefore, we can treat \( \Pi_c \psi_c(x_H, \bar{x}_E) \) as a goal since it is an unnormalized version of both \( p(x_H, \bar{x}_E) \) and \( p(x_H | \bar{x}_E) \) (we just do appropriate normalization).
- To incorporate evidence \( \bar{x}_E \) starting from clique potential \( \psi_c(x_c) \), we just do: \( \psi_c(x_c) \leftarrow \psi_c(x_c) \Pi_{i \in (E \cap c)} \delta(x_i, \bar{x}_i) \)
- To clarify functional notation: \( \psi_c(x) = \psi_c(x_c) \). Also, \( \psi_c(x_H, \bar{x}_E) = \psi_c(x_H \cap c, \bar{x}_E \cap c) \) (i.e., this says what \( \psi \)'s arguments are over).
Inference: outline

• We can treat $\prod_c \psi_c(x_H, \bar{x}_E)$ as a goal since it is an unnormalized version of both $p(x_H, \bar{x}_E)$ and $p(x_H|\bar{x}_E)$ (we just do appropriate normalization). I.e., we have that:

$$p(x_h, \bar{x}_E) \propto p(x_h|\bar{x}_E) \propto \prod_c \psi_c(x_H, \bar{x}_E)$$

• To incorporate evidence $\bar{x}_E$ starting from clique potential $\psi_c(x_c)$, we just do: $\psi_c(x_c) \leftarrow \psi_c(x_c) \prod_{i \in (E \cap c)} \delta(x_i, \bar{x}_i)$

• To clarify functional notation: $\psi_c(x) = \psi_c(x_c)$. Also, $\psi_c(x_H, \bar{x}_E) = \psi_c(x_{H\cap c}, \bar{x}_{E\cap c})$ (i.e., this says what $\psi$’s arguments are over).
Goal: clique potentials as marginals

- Goal: Once process of inference is over, we wish to have the clique potentials be marginals, i.e., goal is for \( \psi_c(x_c) = p(x_c) \), or with evidence \( \bar{x}_E \) we have marginals of the form \( \psi_c(x_{c \cap H}) = p(x_{c \cap H}, \bar{x}_E) \).
- This is sufficient for example to ask many queries, form the joint, or do learning.
- But under (F') property, we can not get that the local potential functions are marginals. I.e., we can’t have \( \psi_c(x) = p(x_c) \) where \( p(x) = \prod_c \psi_c(x) \).
- But if we have a decomposable graph, then we get:

\[
    p(x) = \frac{\prod_{c \in C} p(x_c)}{\prod_{s \in S} p(x_s)} = \frac{\text{prod. of clique marginals}}{\text{prod. of sep marginals}} = \frac{\prod_{c \in C} \psi_c(x_c)}{\prod_{s \in S} \phi_s(x_s)}
\]

With \( \psi_c(x_c) = p(x_c) \) and \( \phi_s(x_s) = p(x_s) \).

- Example: If \( A \rightarrow B \rightarrow C \), then

\[
    p(A, B, C) = p(A)p(B|A)p(C|B) = p(A, B)p(B, C)/p(B).
\]

So the above is a generalization of one of the definitions of conditional independence.
Potentials as marginals

- Again, $\psi_c(x_c)$ and $\phi_s(x_s)$ can be positively scaled while maintaining the above global interpretation. Thus, the above shows that it is possible to have potentials as marginals, but it is not necessarily so.
- The JT algorithm will adjust potential functions, starting from any initial valid configuration (or charge), so that they become actual marginals. valid means that the above global interpretation (i.e., result is $p(x)$) is true.
- But why do we want marginals? First, this is what is minimally necessary for learning. For example, it is easy to calculate prob. of individual nodes.

$$p(x_{c'}) = \sum_{c \setminus c'} \psi_c(x_c) \quad \text{for any } c' \subseteq c$$

- Subsets of nodes in neighboring cliques:

$$p(x_{c_1'}, x_{c_2'}) = p(x_{c_1'} | x_{c_2'}) p(x_{c_2'}) = p(x_{c_1'} | x_s) p(x_{c_2'}) = \frac{p(x_{c_1'}) p(x_{c_2'})}{p(x_s)}$$

$$p(x_{c_1'}) = \sum_{c_1 \setminus c_1'} \psi_{c_1}(x_{c_1})$$
Potentials don’t start as marginals

- So goal is to get potentials as marginals, but there are several reasons why they don’t start out as such.
  1) After introduction of evidence, some clique potentials are unchanged, those with $c \cap E = \emptyset$, and will not be able to produce $p(x_c|\bar{x}_E)$ (but this is what we need).
  2) Not all the evidence $\bar{x}_E$ is present at (probably all) of the clique potentials.
  3) Clique potentials might be initialized by conditional probabilities from directed graph, so might have, say, $\psi_{DEF} = p(F|D,E)$ (not a marginal over $DEF$).
  4) Clique potentials might be specified arbitrarily at first (e.g., MRFs using log-linear models), so potentials are just some unnormalized values, perhaps with some global normalization constant $1/Z$ out in front (which gets cancelled out anyway). In any case, clique potentials wouldn’t be marginals.
Potentials don’t start as marginals

- So, assume we start with:

\[ p(x_H, \bar{x}_E) = \prod_c \psi_c(x_c) = \frac{\prod_c \psi_c(x_H, \bar{x}_E)}{\prod_s \phi_s(x_H, \bar{x}_E)} \]

- If we started from a BN, we moralized, triangulated, and initialized \( \psi_c \) to hold the CPTs for the BN as appropriate, and also initialized \( \phi_s = 1 \).
- If we started from a MRF, again we can choose \( \phi_s = 1 \), and choose any \( \psi_c \) to absorb \( 1/Z \).
- In both cases, evidence is introduced as deltas.
- At the very least, for \( \psi_c \) the clique potentials to be marginals, they must agree on the nodes that they have in common to each other. agree means “be consistent with”, i.e., give the same marginals over intersection. I.e., given cliques \( C_1 \) and \( C_2 \), with \( S = C_1 \cap C_2 \), we must have:

\[
\sum_{x_{C_1 \setminus S}} \psi_{C_1}(x_{C_1}) = \sum_{x_{C_2 \setminus S}} \psi_{C_2}(x_{C_2})
\]
Local consistency

- So far, this is only a necessary (but not a sufficient) condition for marginals. I.e., if they are already marginals, we would certainly have that:

$$\sum_{x_{C_1}} \psi_{C_1}(x_{C_1}) = \sum_{x_{C_1}} p(x_{C_1}) = \sum_{x_{C_2}} p(x_{C_2}) = \sum_{x_{C_2}} \psi_{C_2}(x_{C_2})$$

- Example:

$$\sum_{x_1} p(x_1, x_2) = \sum_{x_3} p(x_2, x_3)$$

- Lets now assume we’ve got two cliques $V$ and $W$ with separator $S = V \cap W$, and potential functions $\psi_V$, $\psi_W$, and $\phi_S$, arranged in a junction tree as follows:

![Diagram](image-url)
A note on notation

- It is common to use the notation $\phi^*_S = \sum_{V \setminus S} \psi_V$. We make sure this is clear.

$\psi_V = \psi_V(x_V)$ represents a clique potential on $x_V$. 

$\phi_S = \phi_S(x_S)$ represents a separator potential on $x_S$.

Also, we have that $V = S \cup (V \setminus S)$, so $V$ contains a part containing $S$ and an “innovation” since $S \subset V$. We can also say $V = S \cup V_S$ where $V_S = V \setminus S$. Thus,

$$\sum_{V \setminus S} \psi_V = \sum_{x_{V \setminus S}} \psi_V(x_V) = \sum_{x_{V \setminus S}} \psi_V(x_{V \setminus S}, x_S) = \phi^*_S(x_S)$$

which is a function only of $x_S$. 
Another note on notation

- *Table multiplication* is also commonly used, and we will see notation like:
  \[ \psi^*_W = \frac{\phi^*_S}{\phi^*_S} \psi_W. \]
  What does this mean?

  Again, we expand out: Let \( W_S = W \setminus S \), so that \( W = S \cup W_S \).

  \[
  \psi_W = \psi_W(x_W) = \psi_W(x_S, x_{W_S}) \quad \phi_S = \phi_S(x_S)
  \]

  \[
  \psi^*_W = \psi^*_W(x_W) = \psi^*_W(x_S, x_{W_S}) \quad \phi^*_S = \phi^*_S(x_S)
  \]

  so we get that

  \[
  \psi^*_W = \psi^*_W(x_S, x_{W_S}) = \frac{\phi^*_S(x_S)}{\phi^*_S(x_S)} \psi_W(x_S, x_{W_S})
  \]
Message Passing

\[ S = V \cap W \neq \emptyset \]

- Suppose we start with things inconsistent. i.e., we have:
  \[ \sum_{V \setminus S} \psi_V \neq \sum_{W \setminus S} \psi_W \quad \text{and} \quad \phi_S = 1 \]

  but we still have that \( p(x_H, \bar{x}_E) = \frac{\psi_V \psi_W}{\phi_S} \)

- The key operation is message passing between cliques via the separators. Multiple messages means exchange of info between clique potentials, which should yield ”consistency”. We have a set of operations:
Message Passing

- Marginalize $V$  
  \[ \phi^*_S = \sum_{V \setminus S} \psi_V \]
  \Rightarrow \text{Updated separator potential}

- Rescale $W$  
  \[ \psi^*_W = \frac{\phi^*_S}{\phi_S} \psi_W \]
  \Rightarrow \text{Update $W$ based on $V$ via $S$.}

- After these operations, the new joint $p(x_H, x_E)$ has not changed. If we define $\psi^*_V = \psi_V$ for convenience, we get:
  \[ \frac{\psi^*_V \psi^*_W}{\phi^*_S} = \frac{\psi_V \psi_W \phi^*_S}{\phi_S \phi^*_S} = \frac{\psi_V \psi_W}{\phi_S} \]
Message Passing

- But while the joint hasn’t changed, we are not yet guaranteed consistency, since:

\[ \sum_{V \setminus S} \psi_V^* = \sum_{V \setminus S} \psi_V = \phi_S^* \neq \sum_{W \setminus S} \psi_W^* = \frac{\phi_S^*}{\phi_S} \sum_{W \setminus S} \psi_W \]

since it may very well be the case that:

\[ \phi_S \neq \sum_{W \setminus S} \psi_W \]

In general, this can be seen as a message passing operator, from \( V \) to \( W \).
Message Passing

- While we don’t yet have consistency, we do have at least one marginal at $\psi^*_W$. This is because we started with:

$$p(x_H, \bar{x}_E) = \frac{\psi_V \psi_W}{\phi_S}$$

and

$$\psi^*_W = \frac{\phi^*_S}{\phi_S} \psi_W = \psi_W \sum_{V \backslash S} \psi_V = \sum_{x_{V \backslash S}} p(x_H, \bar{x}_E) = p(x_{H \cap W}, \bar{x}_E)$$

which is one of the marginals that we desire.
Message Passing

- Marginalize $W$ \[ \phi_{S}^{**} = \sum_{W \setminus S} \psi_{W}^{*} \]
  \[ \Rightarrow \text{Updated separator potential} \]

- Rescale $V$ \[ \psi_{V}^{**} = \frac{\phi_{S}^{**}}{\phi_{S}^{*}} \psi_{V}^{*} \]
  \[ \Rightarrow \text{Update } V \text{ based on } W \text{ via } S. \]

- After these operations, the new joint $p(x_H, x_E)$ has again not changed. If we define $\psi_{W}^{**} = \psi_{W}^{*}$ for convenience, we get:
  \[ \frac{\psi_{V}^{**} \psi_{W}^{**}}{\phi_{S}^{**}} = \frac{\psi_{V} \phi_{S}^{**} \psi_{W} \phi_{S}^{*}}{\phi_{S}^{*} \phi_{S} \phi_{S}^{*}} = \frac{\psi_{V} \psi_{W}}{\phi_{S}} \]
Message Passing

- But consistency is now achieved. In particular, $\psi^*_V$ and $\psi^*_W$ are now consistent since:

$$\sum_{V \setminus S} \psi^*_V = \sum_{V \setminus S} \frac{\phi^*_S}{\phi^*_S} \psi^*_V = \frac{\phi^*_S}{\phi^*_S} \sum_{V \setminus S} \psi^*_V = \frac{\phi^*_S}{\phi^*_S} \phi^*_s = \phi^*_s = \sum_{W \setminus S} \psi^*_W$$

- So we have a forward/backwards message passing scheme on cliques in the graph.

Arrows show dependency graph based on definitions of messages.
Message Passing

- Also, we have the other marginal we asked for at $\psi^*_V$, since:

$$\psi^{**}_V = \frac{\phi^{**}_s}{\phi^*_s} \psi_V = \psi_V \frac{\sum_{W \setminus S} \psi^*_W}{\sum_{V \setminus S} \phi^*_V} = \psi_V \frac{\sum_{W \setminus S} \phi^*_s \psi_W}{\sum_{V \setminus S} \phi^*_V}$$

$$= \psi_V \frac{\sum_{W \setminus S} \psi_W \sum_{V \setminus S} \psi_V}{\sum_{V \setminus S} \psi_V} = \psi_V \sum_{W \setminus S} \psi_W = \sum_{W \setminus S} p(x_H, x_E)$$

$$= p(x_{H \cap V}, x_E)$$

- So, now we have a way both to get the marginals we want and thus achieve local consistency, at least when we have only two cliques in a JT.
Example: A → B → C

- Evidence $C = 1$, $\psi_{AB} = p(A, B)$, $\psi_{BC} = p(C|B)\delta(C, 1)$, $\phi_B = 1$, and $p(A, B, C = 1) = \psi_{AB}\psi_{BC}/\phi_B$.

- Forward:
  $$\phi_B^* = \sum_A p(A, B) = p(B)$$
  $$\psi_{BC}^* = \frac{p(B)}{1} p(C|B)\delta(C, 1) = p(B, C)\delta(C, 1) = p(B, C = 1)$$

- Backwards:
  $$\phi_B^{**} = \sum_C p(B, C)\delta(C, 1) = p(B, C = 1)$$
  $$\psi_{AB}^{**} = \frac{\phi_B^{**}}{\phi_B^*} \psi_{AB}^* = \frac{p(B, C = 1)}{p(B)} p(A, B)$$
  $$= p(A|B)\frac{p(B)}{p(B)} p(B, C = 1) = p(A|B, C = 1)p(B, C = 1) = p(A, B, C = 1)$$
Message Passing

- How can we ensure that this is still correct when we have multiple overlapping cliques, i.e., a tree of cliques?

- Form a junction tree from a triangulated graph $T(G) = (V, E \cup F)$ where $G = (V, E)$ and $F$ are edges necessary to triangulate $G$.
- But how do we make sure that local consistency is not ruined by later message passing steps?
- Ex: Once we send message $V \rightarrow W$ and then $W \rightarrow V$, we know $W$ and $V$ are consistent. If we next send messages $W \rightarrow D_1$ and $D_1 \rightarrow W$, then $W$ & $D_1$ are consistent, but $V$ & $W$ are no longer necessarily consistent.
Message Passing Protocol

- Lets hypothesize that the same message passing scheme that we discovered during tree and polytree inference during Pearl-style inference on trees will work here in the JT tree as well and lets also give it a name, namely the:

**Definition:** Message passing protocol A clique can send a message to a neighboring clique in a JT only after it has received messages from all of its other neighbors.

Ex: $V \rightarrow W$ only after $C_1 \rightarrow V$ and $C_2 \rightarrow V$. 
Theorem: The message passing protocol renders the cliques locally consistent between all pairs of connected cliques in the tree.

Proof:

Suppose \( W \) has received a message from all other neighbors, and is sending a message to \( V \). There are two possible cases:

Case A: \( V \) already sent a message to \( W \) before, so \( V \) already received message from all other neighbors, & message renders the two consistent since neighbor receives any more messages.
Message Passing Protocol & Local Consistency

Case B: V has not yet sent a message to W, so W sends to V & waits. Later, V will have received message from all other neighbors & will send message back to W, but this will contain appropriate update from W.

Another way we can see it: If we abide by the message passing protocol, the potential functions will just be scaled by a constant, and we’ll get back to the same case that we were on page 18 of the notes.

Exercise: show that this is the case, and derive what the constants are.
Hugin propagation (Jensen&Jensen)

Can we find an algorithm that realizes a message scheme that respects the message passing protocol? There are many, the most famous is *Hugin propagation* algorithm.

- Starting with an undirected JT, one clique is arbitrarily designated as the root.
- Children send to parent when they receive messages from all of their children.
- When we reach root, propagate down the tree.
Hugin propagation (Jensen&Jensen)

Procedure $\text{Update}(W \leftarrow V)$
\[
\phi^*_S = \sum_{V \setminus S} \psi_V \\
\psi^*_W = \frac{\phi^*_S}{\phi_S} \psi_W
\]

Procedure $\text{HuginPropagate}(r)$
\[
\text{CollectEvidence}(r) \\
\text{DistributeEvidence}(r)
\]

Procedure $\text{CollectEvidence}(n)$ ($n$ is a node)
for each child $c$ of node $n$
\[
\text{Update}(n, \text{CollectEvidence}(c))
\]
return $(n)$

Procedure $\text{DistributeEvidence}(n)$
for each child $c$ of node $n$
\[
\text{Update}(c, n) \\
\text{DistributeEvidence}(c)
\]
Hugin propagation (Jensen & Jensen)

- Which of the below correspond to Hugin propagation?
- Which of the below obey the message passing protocol?
Theorem: Collect evidence & distribute evidence obey the message passing protocol.

Proof: CollectEvidence (CE) does not send message back to caller (parent) until it has received message from all children (so obeys protocol). After CE is called, all nodes have received messages from all children. Once a node receives message from parent, it can send message to any child. But DistributeEvidence (DE) sends messages to child first, before recursion on that child, so it obeys protocol.

- What happens if we send redundant messages (i.e., we send a message twice)?

\[
\phi^* = \sum_{V \setminus S} \psi_V \quad \phi^* = \sum_{V \setminus S} \psi_V \\
\psi^*_W = \frac{\phi^*_S \psi^*_W}{\phi^*_S} \quad \psi^*_W = \frac{\phi^*_S \psi^*_W}{\phi^*_S} \psi^*_W = \psi^*_W
\]

⇒ so no effect since we divide out previous separator potential at any point in time.
Message Passing Protocol & Local Consistency

- But we need to make sure: 1) that we get marginals, and that 2) that we get global consistency (i.e., all clique potentials agree on their common variables).
**PROBLEM** In this tree of cliques, local consistency would **not** imply global consistency. The reason is that $D$ appears in 2 cliques AED and EDC, but not along the path between them in the tree.

$$\phi_{AE}^* = \sum_D \psi_{AED} \quad \psi_{ABE}^* = \frac{\phi_{AE}^*}{\phi_{AE}} \psi_{ABE}$$

- But here, we have lost AED’s information about $D$, we can’t pass this along to EDC. So we wouldn’t get global consistency since info about $D$ is lost (the two ”marginals” on AED an EDC would produce different results for $p(D)$). Also, we wouldn’t even get marginals at all!!
Message Passing Protocol & Local Consistency

• Real issue: We don’t have a JT, and the reason why is that we didn’t have a triangulated graph to begin with.
• Recall crucial property of JTs: For every pair of cliques $V$ and $W$, all cliques on the necessarily unique path between $V$ and $W$ will contain $V \cap W$. This property will mean that achieving pairwise consistency will guarantee global consistency.

**Theorem:** Suppose we have a true junction tree (JT). Let $p(x_H, \bar{x}_E)$ be represented w.r.t. this JT as follows:

$$p(x_H, \bar{x}_E) = \frac{\prod_c \psi_c(x_c)}{\prod_s \phi_s(x_s)}$$

After the JT algorithm finishes (using either Hugin propagation, or any other procedure that obeys MPP), then we will have that $\forall c \ \forall s$:

$$\psi_c = p(x_c, \bar{x}_E) \quad \phi_s = p(x_s, \bar{x}_E)$$
Reminders

• no class next week
• Midterm on web page Friday evening, due next Friday evening. Start early!!